UNIVERSITY OF EXETER
Q-STEP CENTRE

Q-Step
A step-change in quantitative social science skills
Funded by the Nuffield Foundation, ESRC and HEFCE

Probabilistic logic and statistical inference

Part 2: Continuous variables

Lewys Brace
l.brace@Exeter.ac.uk

Continuous variables

- In the last lecture, we looked at discrete variables; coin flips and dice rolls.
- But what about continuous variables?
- These are variables whose quantities can take any value, not just discrete ones; i.e. a car can travel at 62.6 miles per hour.
- Continuous variables also have probability distributions.

An example

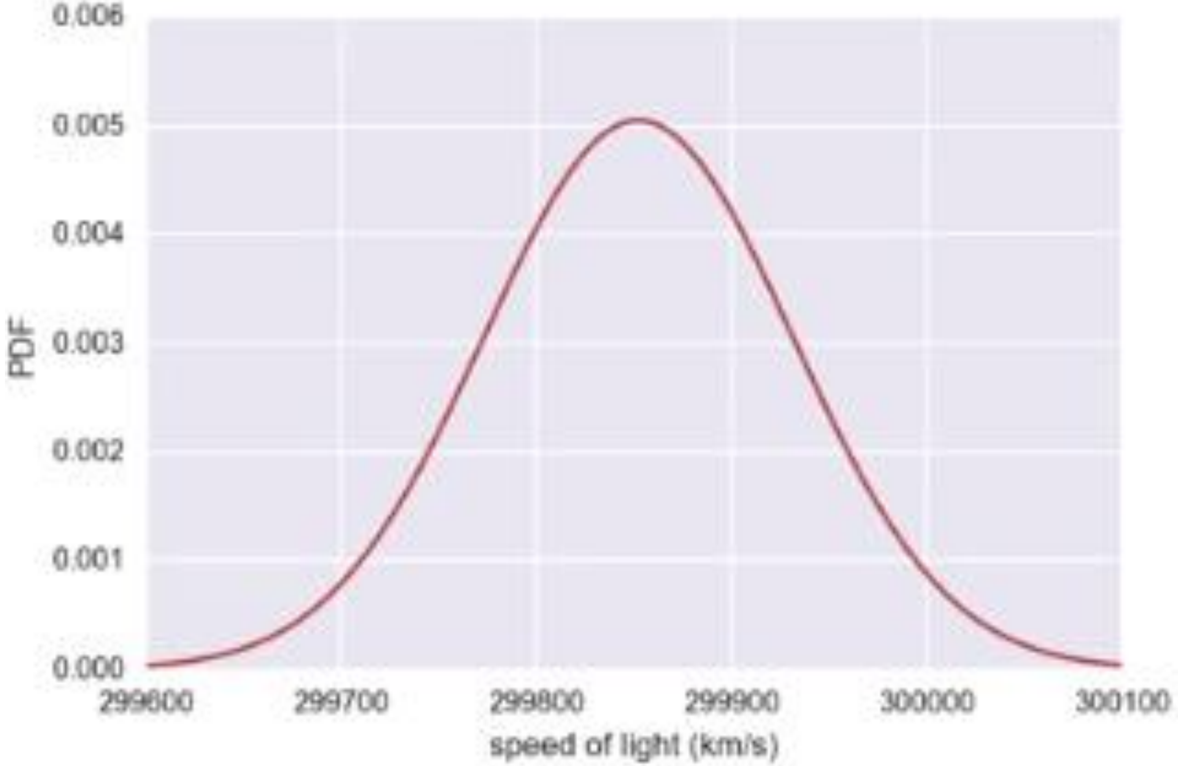
- In 1879, Albert Michelson performed 100 measurements of the speed of light and air.
- Each measure had an associated error of some kind; i.e. temperature, alignment of optics, etc, changed from experiment to experiment.
- As a result, any fractional value of the measured speed of light is possible.
- It is therefore apt to describe the result with a continuous probability distribution.

- Which probability distribution describe these data?
- The normal distribution?
- In order to understand the normal distribution, lets look at its *probability density function*.

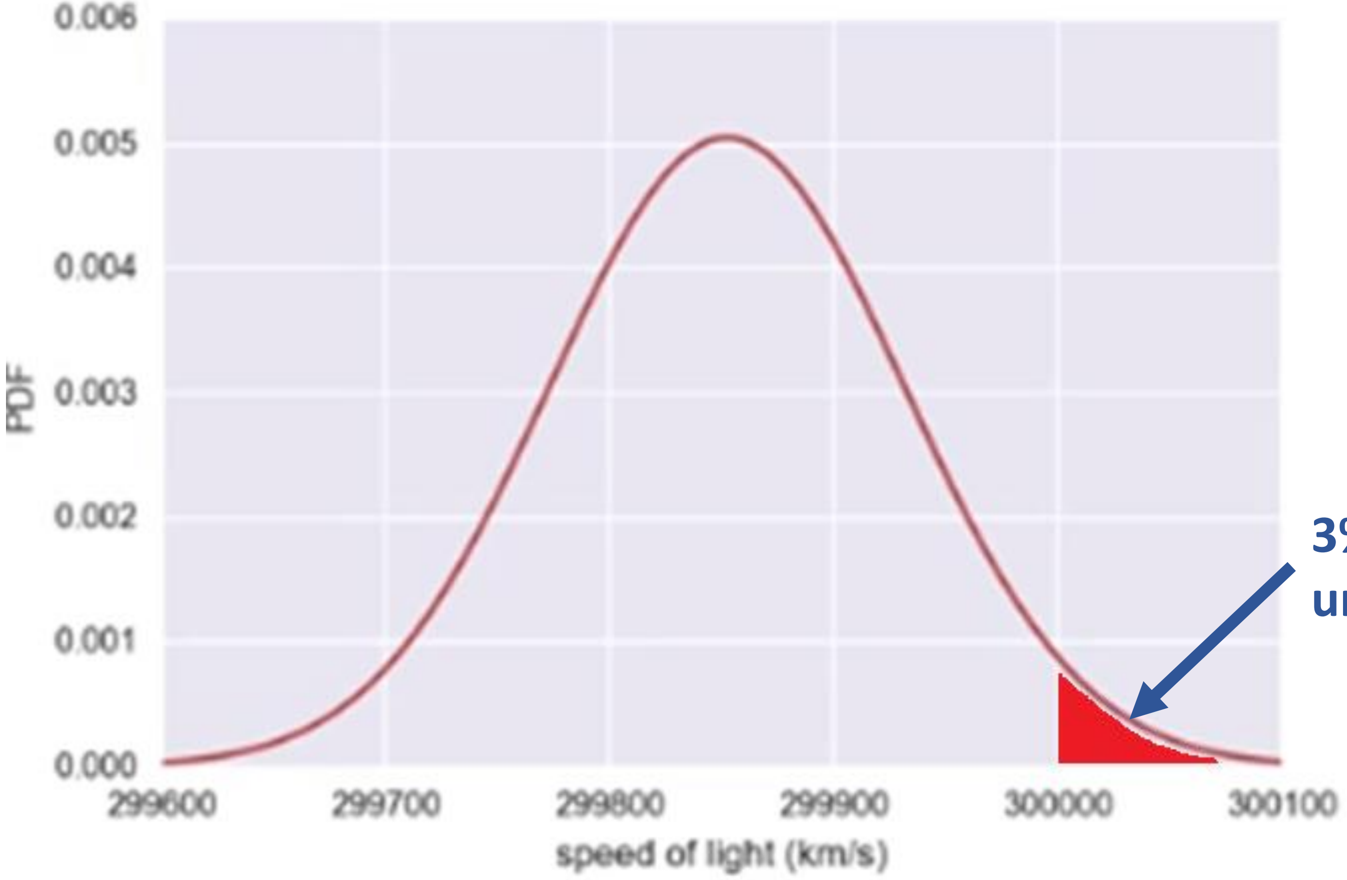
Probability density function (PDF)

- This is the continuous analogue of the PMF from the last lecture.
- It is the mathematical description of the relative likelihood of observing a value of a continuous variable.

- The probability of observing a single value of the speed of light doesn’t make sense; because there is an infinity of numbers, say between 299600 and 300100 km/s.
- Instead, the areas under the PDF give probability.
- So…

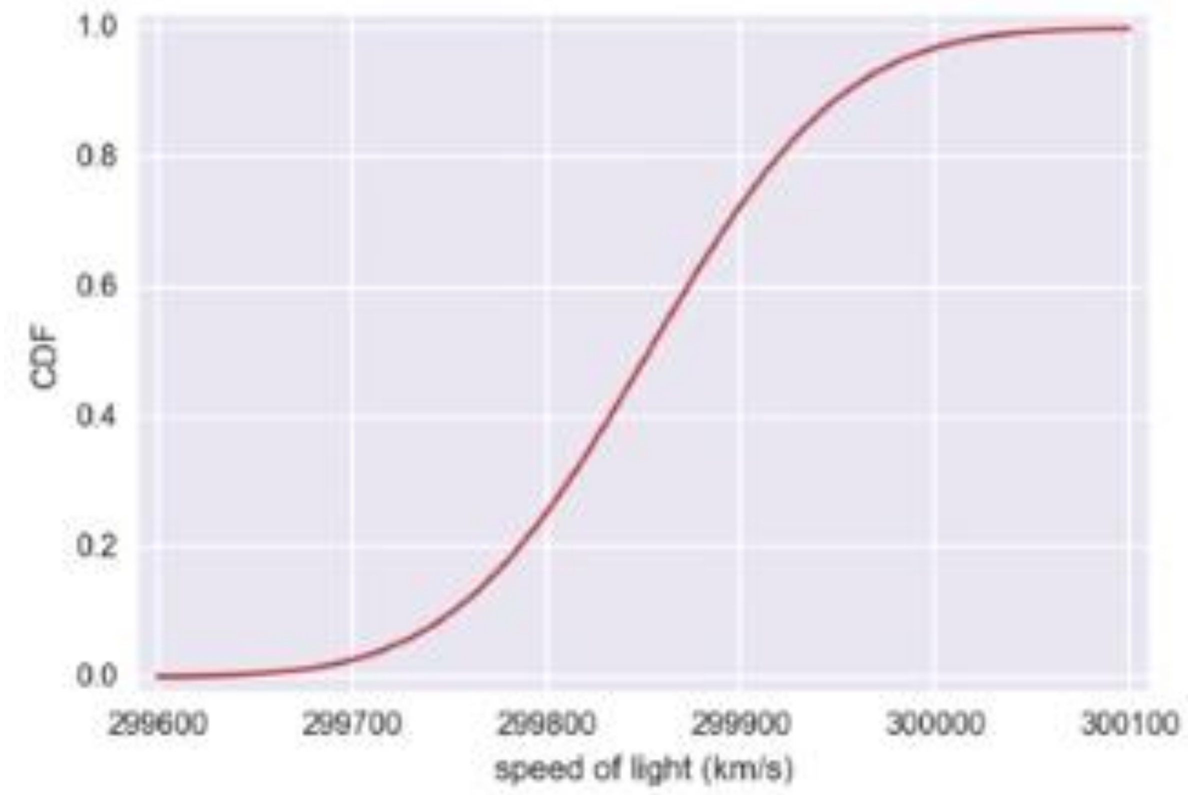


- The probability of measuring the speed of light to be greater than 300000 km/s is an area under the normal curve.
- Parameterising the curve based on Michelson’s experiments, there is about a 3% chance of this happening.

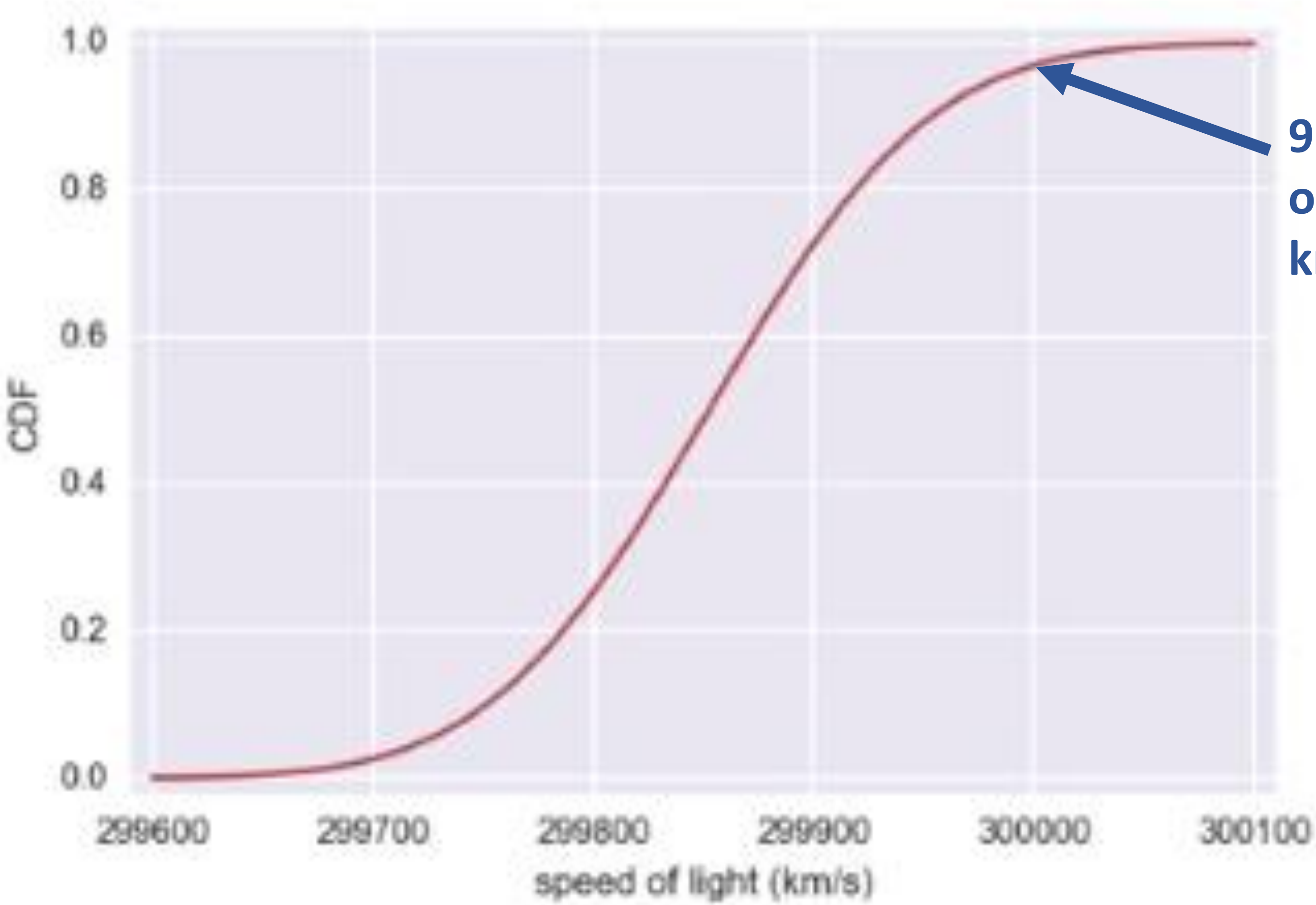


3% of total area under PDF

- In doing this calculation, we were really just looking at the cumulative distribution function (CDF) of the normal distribution.
- Remember, the CDF gives the probability give the probability that the measure speed of light will be less than the value on the X-axis.



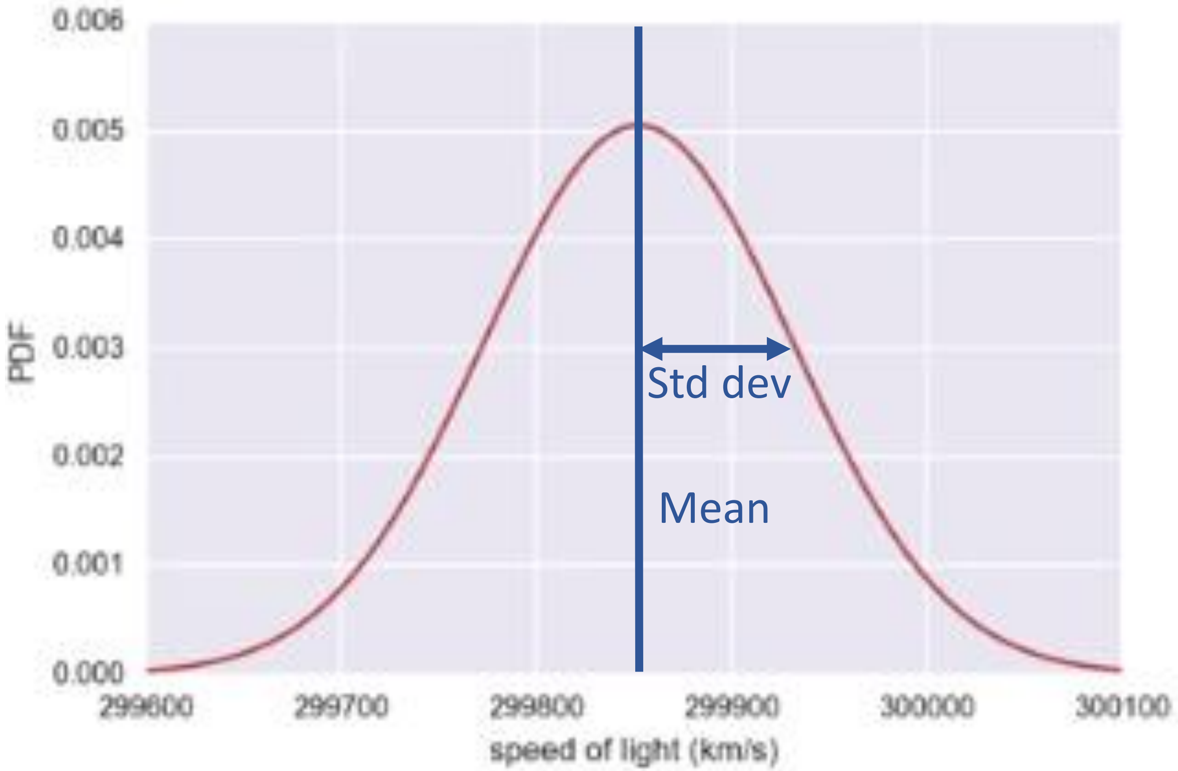
- So, reading of the graph at 300,000 km/s, we see that there is a 97% chance that a speed of light measurement will be less than that; so there is a 3% chance that it is greater.



97% chance speed of light is <300,000 km/s

The normal (Gaussian) distribution

- The normal distribution describes a continuous variable whose PDF has a single symmetric peak.
- It is parameterised by two parameters.
- The mean determines where the center of the peak is.
- The standard deviation is how wide the peak is, or how spread out the data are.

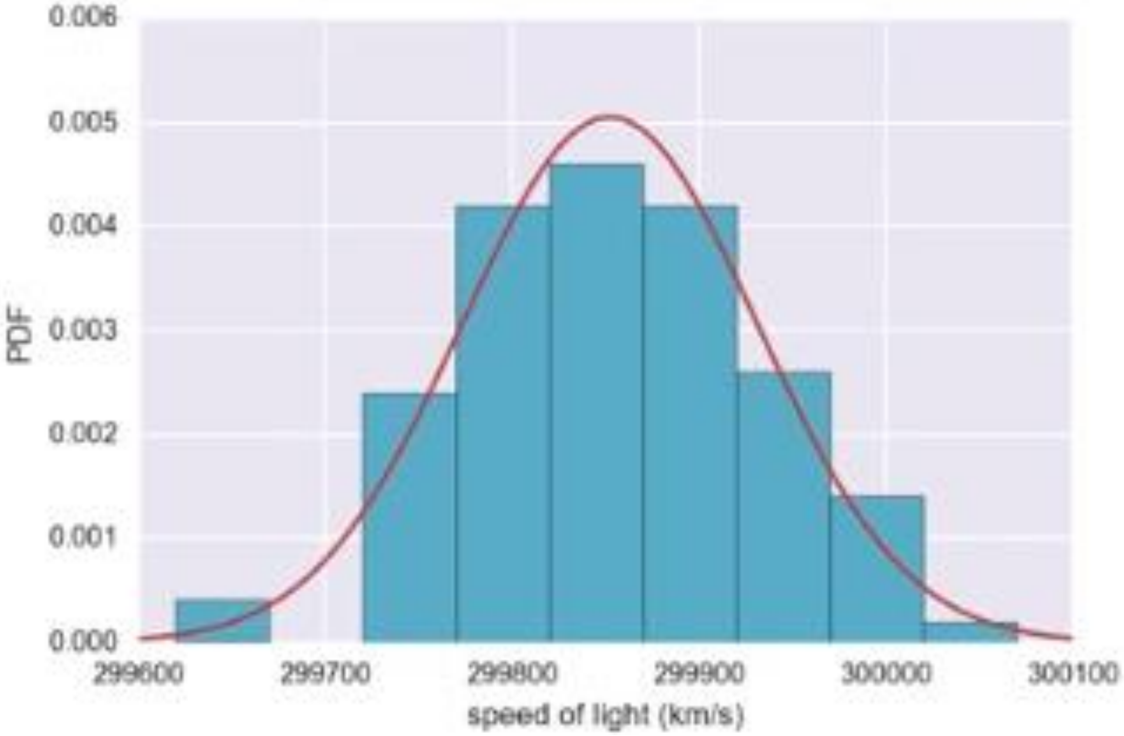


An important note

Parameter		Calculated from data
Mean of a normal distribution	≠	Mean computer from data
Standard deviation of a normal distribution	≠	Standard deviation computer from data

Comparing data to a Normal PDF

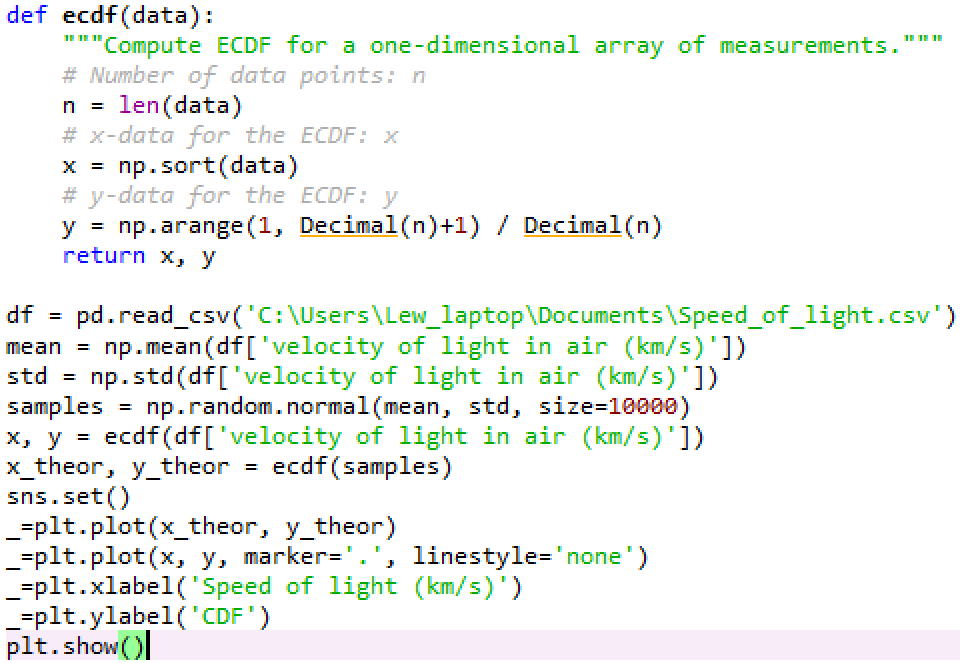
- Adding a histogram of the measurements, we see that the measured speed of light and air appears to be normally distributed.



- However, comparing the histogram to the PDF suffers from binning bias.
- It is therefore better to compare the ECDF of the data to the theoretical CDF of the normal distribution.
- We can again is Numpy to draw samples, and then compute the CDF.

Checking Normality of Michelson data

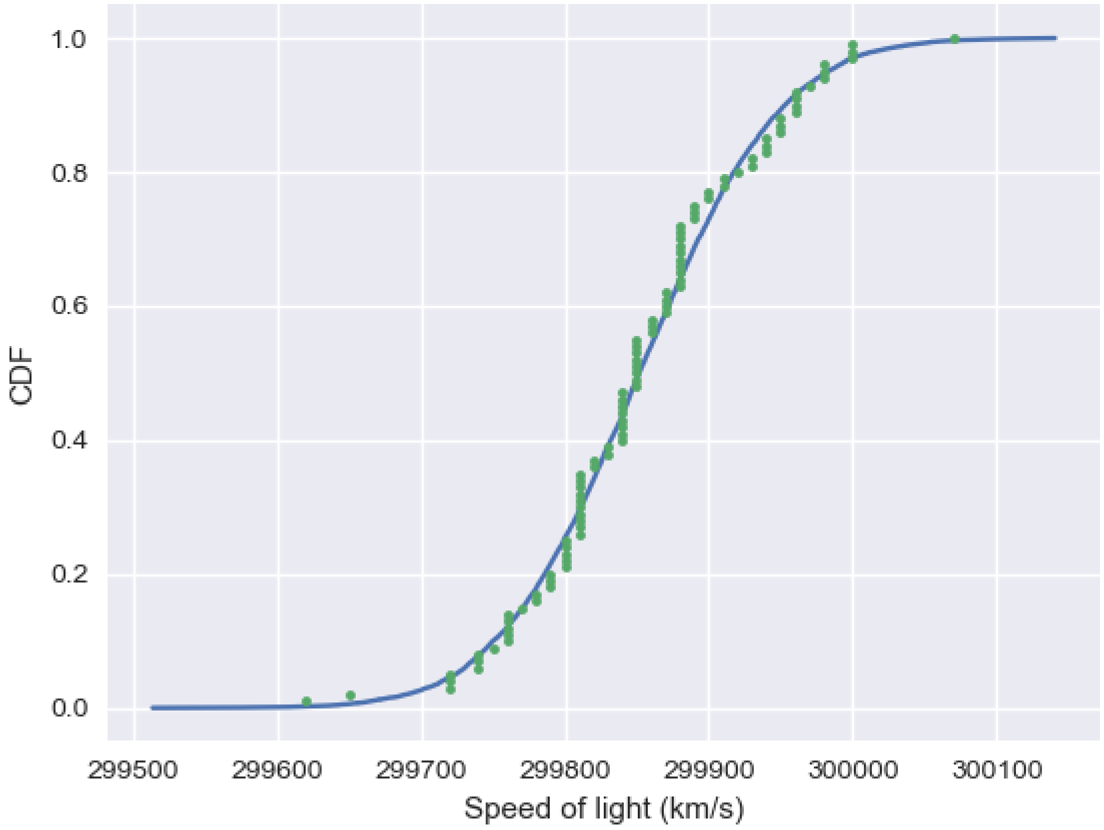
In:



```
def ecdf(data):
    """Compute ECDF for a one-dimensional array of measurements."""
    # Number of data points: n
    n = len(data)
    # x-data for the ECDF: x
    x = np.sort(data)
    # y-data for the ECDF: y
    y = np.arange(1, Decimal(n)+1) / Decimal(n)
    return x, y

df = pd.read_csv('C:\Users\Lew_laptop\Documents\Speed_of_light.csv')
mean = np.mean(df['velocity of light in air (km/s)'])
std = np.std(df['velocity of light in air (km/s)'])
samples = np.random.normal(mean, std, size=10000)
x, y = ecdf(df['velocity of light in air (km/s)'])
x_theor, y_theor = ecdf(samples)
sns.set()
_=plt.plot(x_theor, y_theor)
_=plt.plot(x, y, marker='.', linestyle='none')
_=plt.xlabel('Speed of light (km/s)')
_=plt.ylabel('CDF')
plt.show()
```

Out:



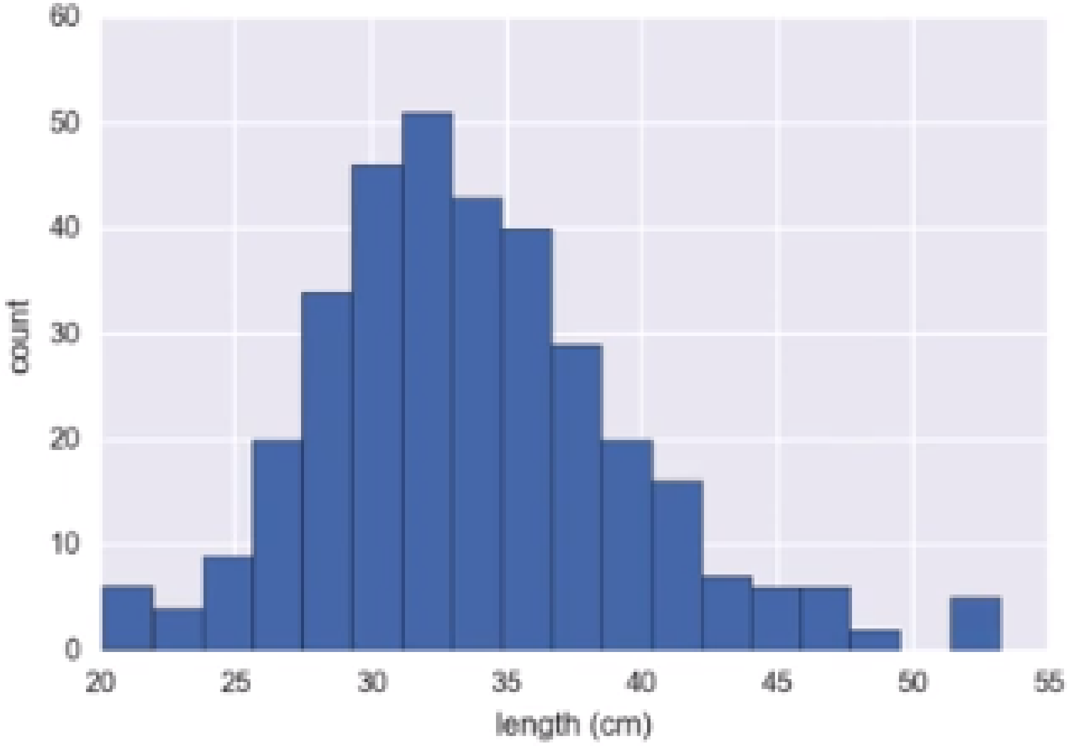
- With the absence of binning bias, it is easy to see that Michelson’s data is indeed normally distributed.

Normal distribution: Properties and warnings

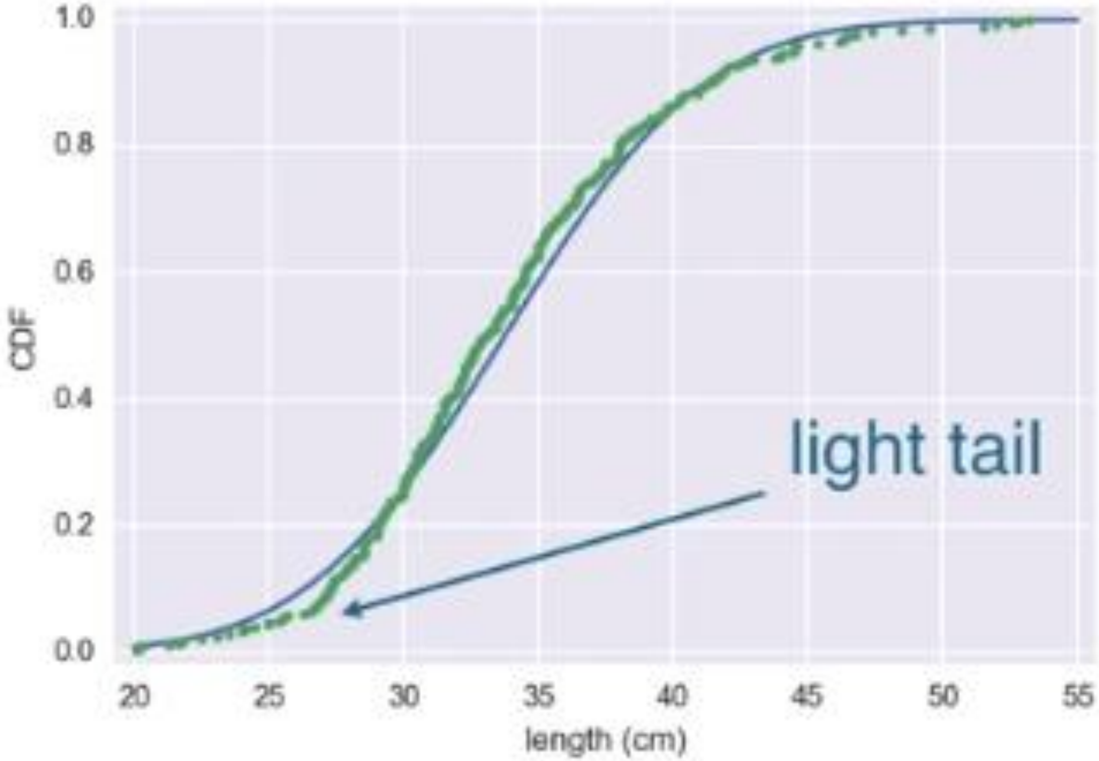
- In practice, the normal distribution is used to describe most symmetric, peaked, data that you will encounter.
- For many of the statistical procedures that you have heard of, normality assumptions about the data are present.
- However, there are important caveats for the normal distribution, and we need to be careful using it.

Things are not always normal!

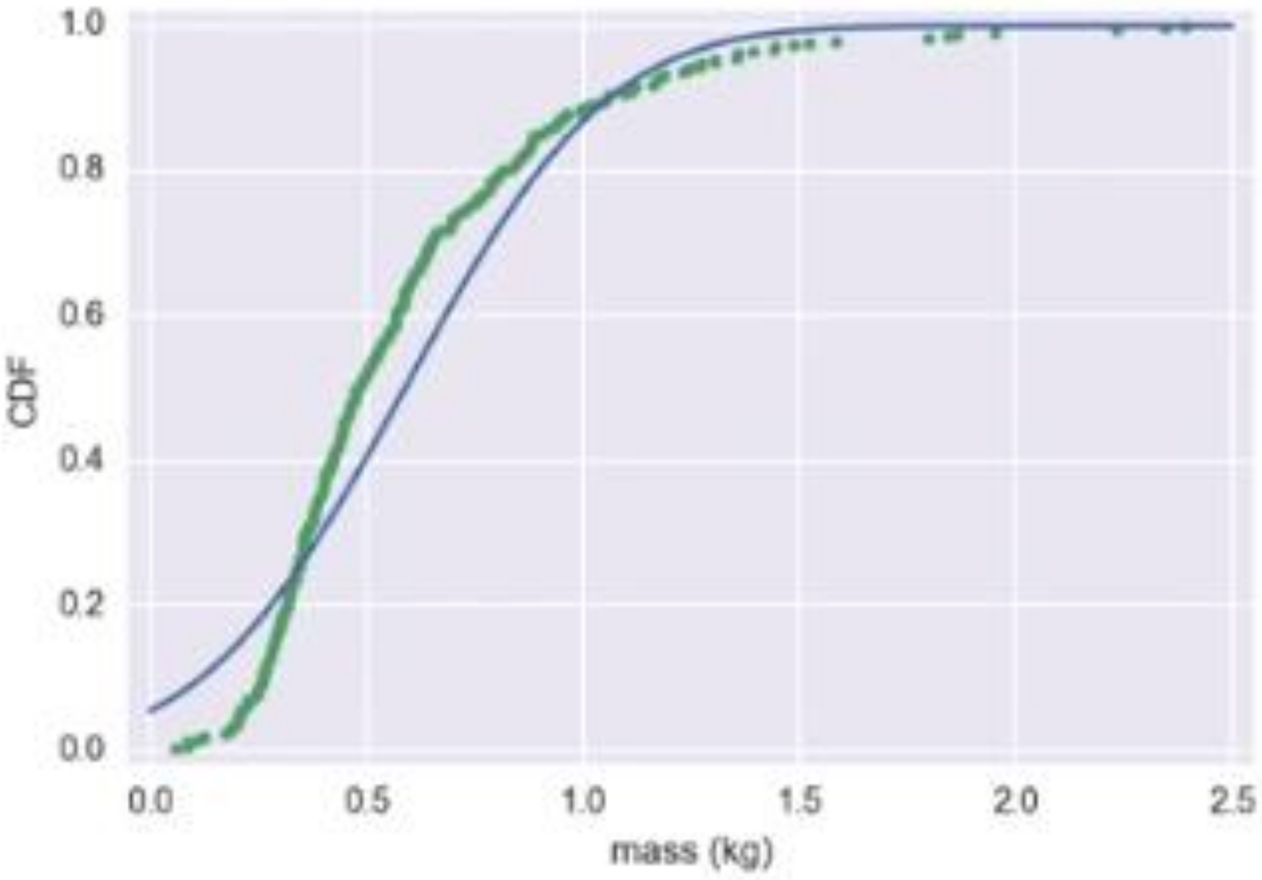
- First, sometimes things you think are normal distributed are not.
- As an example, lets look at the large mouth bass found in the lakes of Massachusetts, USA.
- These were measured both in 1994 and 1995.
- The histogram of fish lengths indicates that the data is normally distributed.



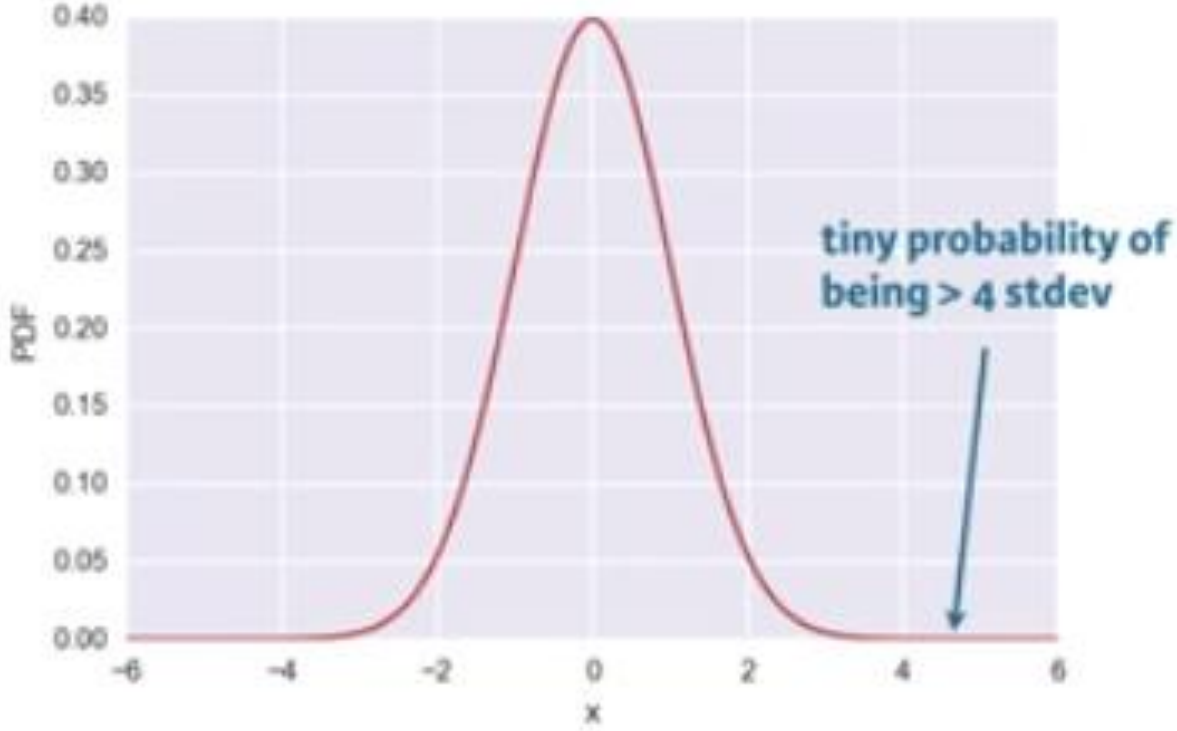
- The same is true if we look at the ECDF overlaid with a theoretical normal CDF.
- Although, there are some systematic differences, particularly on the left tail.
- So, this may not be quite a perfect normal distribution, but we wouldn’t be making too much of a big error by treating it as so.



- What a bout the mass of the bass?
- If the length of the bass is normally distributed, then it follows that the mass will also be normally distributed, right?
- Not so!
- Thus, our initial assumption was wrong



- Another important issue to keep in mind when using the normal distribution is the lightness of its tails.
- Looking at the normal distribution, the chances of being more than 4 standard deviations from the mean are very small.
- Therefore, when modelling data that is normally distributed, outliers are incredibly unlikely.

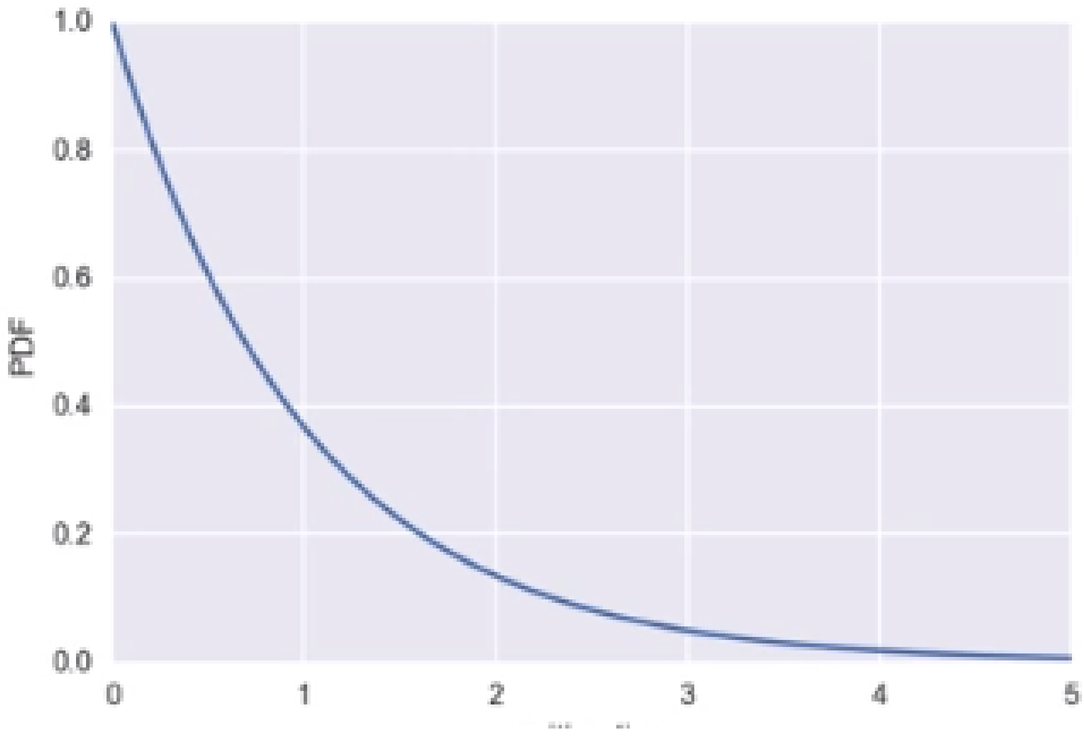


- However, as we know, real data sets often have extreme values.
- In such cases, the normal distribution may not be the best description of the data.

Remember to think about the assumptions that go into your analysis!

The exponential distribution

- Just as there are many names for different discrete distributions, there are also many continuous distributions.
- One is the exponential distribution.

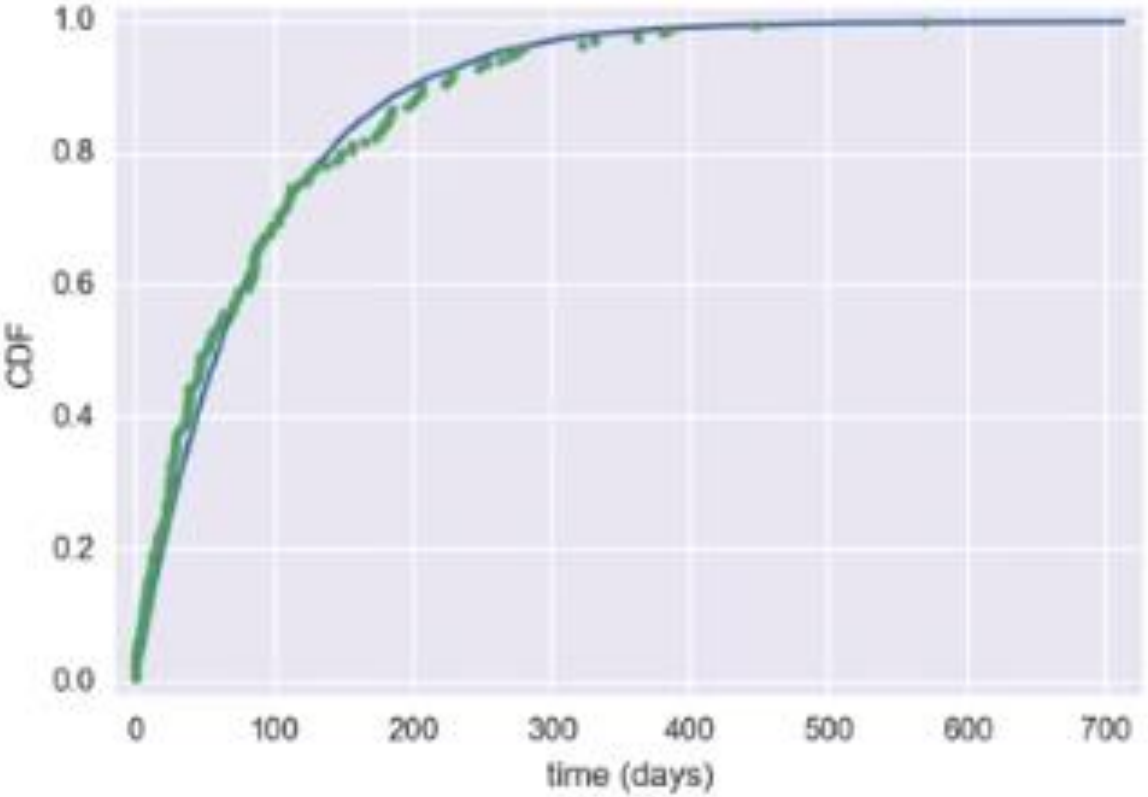


An example

- Lets look at all incidents involving nuclear power since 1974.
- We may expect incidents to be well modelled by a Poisson process; i.e. the timing of one event is independent of all others.
- So, the time between incidents should be exponentially distributed.
- We can compute and plot the CDF we would expect, based upon mean time between incidents, and overlay that with ECDF of the real data in the same manner as we have done previously….

Exponential inter-incident times

- We see that it is close to being exponentially distributed.
- This indicates that nuclear incidents can indeed be modelled as a Poisson process.



Distributions

- The normal and exponential distributions are just two examples of many continuous distributions.
- Importantly, in many cases, you can just simulate your story to get the CDF.
- Remember, you have computers at your disposal, if you can simulate a story, you

Any questions?